A Short Note on Surveillance Testing

FdT, CG, & VK on behalf of LMH Governing Body

13/09/2020

This document outlines some simple strategies to assess the prevalence of coronavirus when there is insufficient capacity to test everyone in the population of interest. The problem of trying to determine whether a given individual has coronavirus is fundamentally different from the problem of estimating the prevalence of infection. The high test specificity and sensitivity along with rapid turnaround that are crucial for individual medical diagnosis are not absolutely essential for surveillance testing. The purpose of this document is to suggest that only relatively modest resources may be sufficient to obtain valuable information regarding the spread and prevalence of coronavirus in the Oxford population. We view this issue as separate from and parallel to the the "Early Alert Testing" service. The ideas we discuss are neither novel nor original, and we do not propose a detailed plan for surveillance testing. Our goals are merely to explain randomized and pooled testing strategies to a non-specialist audience, and to calculate the approximate number of tests required to provide useful information. For simplicity we begin by abstracting away the problem of false negatives and false positives. The final section discusses how these affect our analysis.

The Bottom Line

Testing a random sample of 1400 students each week would provide useful estimates of the prevalence and spread of coronavirus at Oxford, allowing us to evaluate the effectiveness of the university's COVID-19 Student Responsibility Agreement and guide potential policy changes. Using a simple pooling strategy, this would require approximately 224 laboratory tests per week, i.e. 32 per day. Because capacity cannot be diverted from the Early Alert Testing service, the university would need to obtain a total of roughly 2000 laboratory tests from another source over the course of MT. These need not have a rapid turnaround time: a delay of several days would still be useful to assess the spread of infection. Pooled testing reduces the number of laboratory tests required but does not affect the number of samples that need to be collected. To lessen the logistical burden of collecting 200 samples each day, individuals could be asked to self-swab while being supervised by a healthcare professional.[1]

[1]This is the procedure used by the ONS in their Coronavirus Infection Survey pilot. A recent study suggests that this procedure is no less accurate than having healthcare workers conduct the swabs directly: https://www.medrxiv.org/content/10.1101/2020.04.11.20062372v1.

Randomized Testing

As a rough approximation, suppose that there are 20,000 undergraduate and taught masters students in residence at Oxford during MT. An unknown number C of them have the coronavirus. With a perfect test, we could learn C by testing all 20,000 students. But what if we only have the capacity to carry out n tests, where n is much smaller than 20,000? In this case we cannot learn C exactly, but by employing *randomized testing* we can produce an estimate that is sufficiently accurate to guide our mitigation policies. The value of randomized testing is widely acknowledged. It underlies both the Coronavirus Infection Survey pilot conducted by the ONS, and the REACT-1 survey conducted by the Department of Health and Social Care. Both of these surveys are extremely valuable, but as they aim to be nationally representative, neither employs a sufficiently large sample size to track local trends. As such we propose that randomized surveillance testing be carried out within the university.

The most basic form of randomized testing relies on a *simple random sample*. To use this procedure, we place all 20,000 student id numbers into a hat, mix them up thoroughly, and then draw out n at random. These n students are sent for a coronavirus test. We then use the number of students in our sample who have coronavirus to calculate an estimate E of the number students in the population who have coronavirus. For example, if one student in a sample of 2000 tests positive, then we would estimate that there are there are 10 cases in the University as a whole.

Because students are chosen at random, the estimated number of cases E may not equal the true number of cases C. Purely by chance, we might draw a sample of students in which no one is infected. In this case E would equal zero: an underestimate of C. We could also be unlucky in the opposite direction and, purely by chance, draw a sample that contains everyone at Oxford who is infected. In this case E would be an overestimate of C. Both of these possibilities, however, are remote: E is highly likely to be close to C. With an appropriate sample size, an estimate based on random testing is reliable enough to tell us with high confidence whether the number of cases is, say, greater than 50 or fewer than 20. It can also be used to determine whether prevalence is increasing and, if so, how quickly.

Relative to the Early Alert Testing system, which tests only symptomatic individuals, a crucial advantage of randomized testing is that it will also detect pre-symptomatic and asymptomatic cases. This is important because, as explained in the SAGE report of 3rd September, "asymptomatic transmission is a key risk in university settings." Indeed, a recent study cited in the SAGE report finds that only 18% of those aged 0–19 and 22% of those aged 20–39 who are infected with coronavirus show symptoms of the disease.[2] Even if these figures are substantial overestimates, it is likely that a considerable fraction of infections among our students will go undetected by the Early Alert Testing system. For this reason, even if the turn-around times for randomized tests were somewhat slower, they could still provide more timely information about the spread of infection.

The key question in designing a randomized testing protocol is how large a sample size to use. Fortunately, there are several simple ways to provide an approximate answer. One is to

[2] https://arxiv.org/abs/2006.08471

calculate how many students we would need to sample *on average* before we detect our first positive result. If there are C total cases out of 20,000 students in the university then the answer is approximately $20{,}000/C$.[3] The following table presents results for a range of values for C. As we see from the table, the more prevalent the infection, the fewer tests we need before we detect a positive. For purposes of comparison, we can say with high confidence based on recent results from REACT-1 that between 0.16% and 0.41% of individuals in the 18-24 age group had the coronavirus at the start of September: roughly 30-80 out of 20,000.

Table 1: Expected number of samples before the first positive result under random sampling without replacement from a population of size 20,000.

#Cases (C)	Prevalence (%)	Expected #Samples
20	0.10	952
30	0.15	645
40	0.20	488
50	0.25	392
60	0.30	328
70	0.35	282
80	0.40	247
90	0.45	220
100	0.50	198

A second way to address the question of how many samples are needed is by calculating *how close* the estimated number of cases E will likely be to the true number of cases C for different values of n. Suppose that there are 60 cases within the university. This would represent a slight uptick from the most recent REACT-1 estimate of coronavirus prevalence among 18-24 year olds. With a sample size of $n = 1000$ the estimated number of cases E will fall in the range (20, 100) with more than 80% probability. With a sample size of $n = 1500$ this range would narrow to (26, 94). The figure overleaf presents analogous results for sample sizes of 500 and 2000.[4]

A third way of deciding how many samples are needed is by calculating the chance that we will be able to reliably distinguish between two different levels of coronavirus prevalence given a particular sample size. Suppose, for example, that we initially believed there to be 50 cases in the university, corresponding to the current estimated prevalence among the university-aged population from REACT-1. If cases increased to 100, we would have approximately a 68% chance of detecting this, with high confidence, based on a sample 1400 students.[5] If they increased to 200, we would have a 99% chance of detecting it. Intuitively, if cases were

[3] The exact value is $(N+1)/(C+1)$ for a population of size N containing C positives.

[4] These intervals are calculated from the quantiles of a Hypergeometric(C, $N-C$, n) distribution where n is the sample size and C is the number of positives in a population of size N.

[5] The calculations in this paragraph are based on a power calculation for a test of the null hypothesis $H_0\colon p \leq 0.0025$ against the one-sided alternative $H_1\colon p > 0.0025$ with level $\alpha = 0.1$ using the score test statistic $T_n = (\widehat{p} - p_0)/SE(p_0)$ where $SE(p_0)$ is the theoretical standard error under the null with a finite-

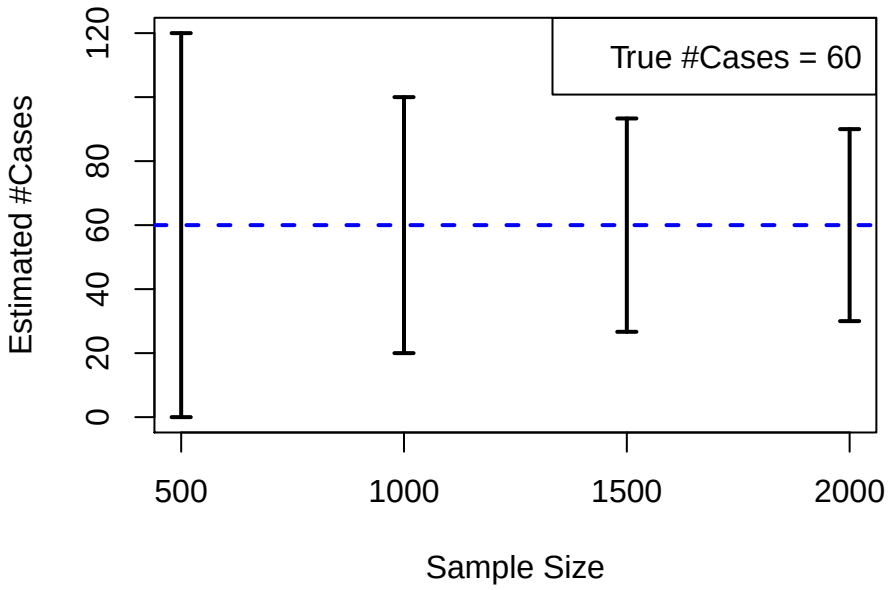


Figure 1: This figure shows how close the estimated number of cases will likely be to the true number of cases for different sample sizes. At a given sample size, the estimated number of cases as a greater than 80% chance of falling in the specified interval.

doubling weekly, then a sample of roughly 200 students per day would be likely to detect an increase after one week, and nearly certain to do so after two weeks.

With 200 tests per day, we would have a roughly even chance of detecting this, with high confidence, in one week, and be almost certain to detect it within two. (FOOTNOTE)

For example, suppose that we can carry out 200 random tests per day. Then within five days we can with high confidence distinguish between a situation in which there are 100 cases in the university as opposed to 20. (FOOTNOTE.)

As another example,

Approximately 200 tests per day would already be useful for answering basic population-level questions concerning prevalence and spread. This would amount to around half as many tests over the entire term as would be required for a *single* round of universal testing.

A simple random sampling procedure is the easiest to explain and to use for back-of-the-envelope calculations, but there are several ways that it could be refined to increase the accuracy of our estimates. It is likely that coronavirus cases will tend to "cluster" within colleges, since the infection is most easily spread through frequent, close contact. If this is so, then a sampling procedure that allocates a *fixed* number of tests to each college (proportional to its size) and draws a simple random sample *within* each college will be more efficient than a simple random sample taken at the university level. This is called a *stratified* random sample. Another possibility is to employ *adaptive sampling*, in which additional tests are

population correction, i.e. $SE(p_0) = \sqrt{\kappa p_0(1-p_0)/n}$ where $\kappa = (N-n)/(N-1)$ and N is the population size.

allocated to colleges in which positive results are detected in a first batch of tests. While the analysis of such a design is more complicated mathematically, it makes more efficient use of a limited number of tests.

Pooled Testing

Pooled testing is a procedure that can be helpful in situations where laboratory capacity, rather than the ability to collect test samples, is the limiting factor in expanding testing. In this procedure, a single laboratory test is carried out on the *combined* samples of a group of individuals. For a sufficiently sensitive test, a negative result for the pooled sample indicates that no one in the group is infected. A positive test indicates that at least one individual in the group is infected, and it is then necessary to test the individual samples to pinpoint the infected individual(s). When the infection is sufficiently rare, pooled testing is highly efficient because virtually all pooled samples will test negative. (Footnote: if you're mainly interested in isolating rather than prevalence, don't even need to re-test! So then just divide by batch size.) For example, if $C = 20$ then pooling tests by household into groups of six requires approximately 3500 laboratory tests to achieve the equivalent of universal testing. (Footnote with calculation, explain pessimistic assumptions). This strategy has in fact been adopted by the University of Cambridge. In the absence of sufficient laboratory capacity, pooled testing could be combined with randomized testing.

```
p1 <- 20/20000
p2 <- 100/20000
k <- 6
c(1 / k + 1 - (1 - p1)^k, 1 / k + 1 - (1 - p2)^k)
```

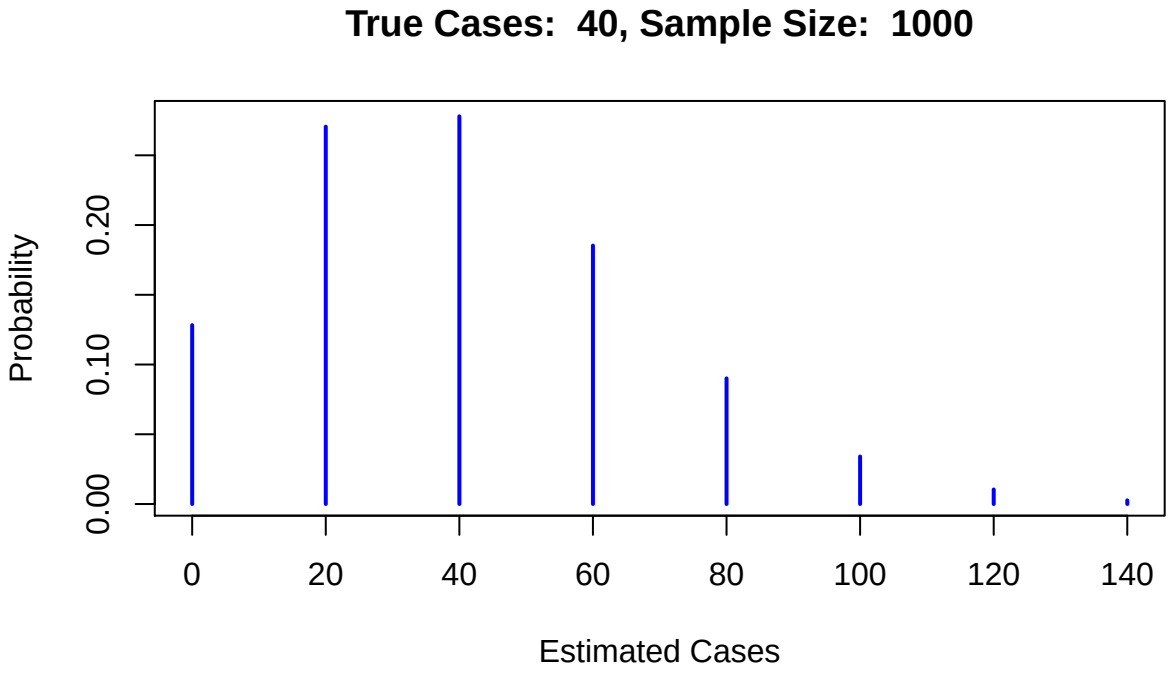
```
## [1] 0.1726517 0.1962942
```

What about false positives and negatives?

Thus far we have abstracted away the problem of false positives and false negatives. As our primary goal is to estimate prevalence and spread rather than to diagnose particular individuals, this is not a major concern. First, if false positive and false negative rates are constant, we can still infer the spread of the infection by examining *changes* in estimated prevalence. Second, recent research has proposed estimates of these rates that can be used to correct estimates of population prevalence.

This latter point is best understood using an example. Suppose that there are 40 coronavirus cases among 20,000 students. This amounts to a prevalence of 0.2%, which would be equivalent to two doublings of the ONS estimate of 0.05% from late August. Accounting for the fact that cases are currently more common among younger age groups, this may be a reasonable approximation to the state of play near the beginning of MT. The following figure shows the different values that the estimated number of cases $\widehat{C}$ could take on, along with their probabilities. In statistical parlance, this is called the "sampling distribution" of $\widehat{C}$. We

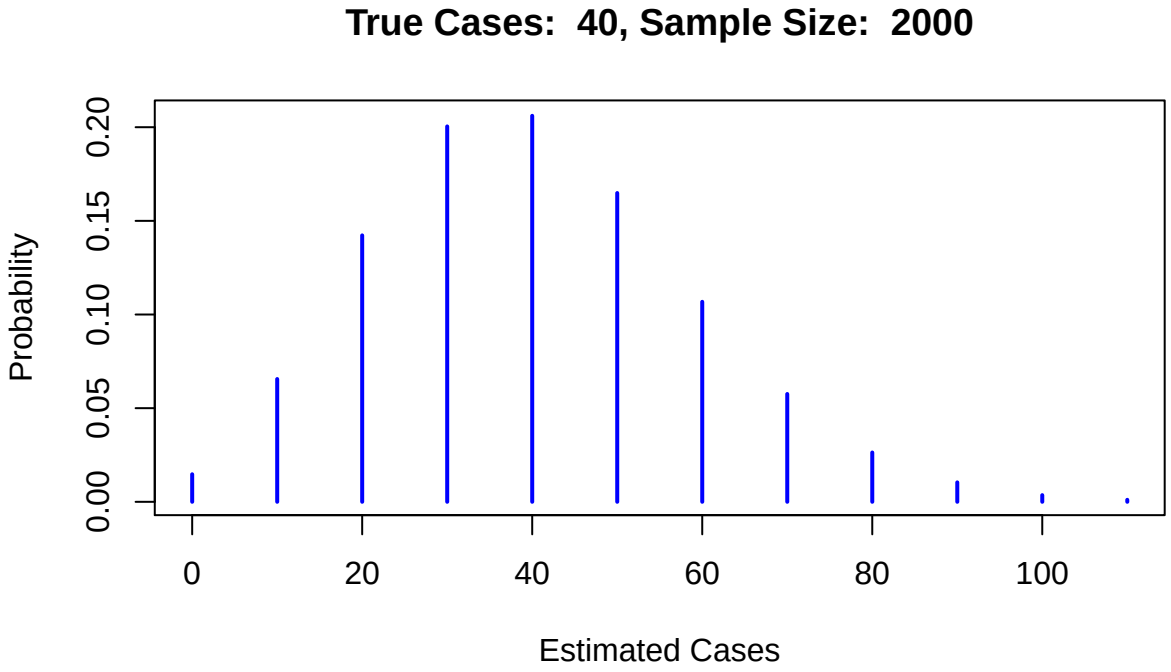
can calculate these probabilities *exactly* because the randomness in $\widehat{C}$ is entirely under our control: it comes from the fact that students are randomly chosen to be tested.




```
d_C_hat <- function(x, n, C, N = 20000) {
  dhyper(n * x / N, C, N - C, n)
}
```

From the figure, we see that $\widehat{C}$ there is a 28% chance that our estimate $\widehat{C}$ will equal 40, the true number of cases. Moreover, there is a 73% chance that it will be between 20 and 60. There is, of course a chance that our sample will not contain any positives: $\widehat{C}$ equals zero with 13% probability.

For a given number of true cases in the population, increasing the sample size makes it more likely that $\widehat{C}$ will be close to C. For $n = 2000$ the sampling distribution of $\widehat{C}$ becomes more "bell-shaped." With this increased sample size, there is now a 82% chance that it will be between 20 and 60. The chance that we will fail to detect any cases falls to 1%.



For a fixed sample size, the probability of failing to detect any cases also falls as the number of true cases rises. Suppose, for example, that the number of cases in the population were to double from 40 to 80. Then, with a sample size of 1000, the probability that $\widehat{C} = 0$ would fall from 13% to 2%. As a rough guide, a sample size between one and two thousand tests per week should be sufficiently informative to detect cases when the prevalence is 0.2% or above.